关联数据在学术资源网相似文献发现中的应用研究*

赵夷平 毕 强
(吉林大学管理学院 长春 130022)

摘要: 【**目的**】利用关联数据的机器可读、语义表示、关联描述和网络资源属性的优势，弥补学术资源网信息组织的不足，为相似文献发现提供支持。【**方法**】采用潜在语义分析方法计算学术资源网发布的文献的总体相似度，通过层次聚类方法确定相似度阈值进行相似度筛选，生成文档关系矩阵，在此基础上利用动态文档技术构造学术资源网关联数据以支持关联文献语义检索。【**结果**】初步实现具有相似文献查询功能的学术资源网关联数据，用于便捷地获得与任何一篇文献高度相关的文献，有助于高效地发现相似文献。【**局限**】仅从统计学角度实现学术资源网中相似文献的发现，对于利用文档集知识体系、语义内涵和组织方式等进行深度的相似文献发现有待进一步研究。【**结论**】潜在语义分析方法计算文献相似度可有效发现相似文档，将相似文献关联记录在关联数据中，支持语义检索获得精确的相似文献，并能够大幅缩减实时相似性计算的延迟。

关键词: 关联数据 潜在语义分析 学术资源网 相似度
分类号: G354

1 引 言

学术资源网是供从事某个领域科研工作的学者们发表学术见解和学术成果、交流学术思想的网络空间，它蕴藏着极为丰富的学科信息资源。如科学网(www.sciencenet.cn)、统计之都(cos.name)、小木虫(emuch.net)、中国万维网联盟(w3china.org)等。这些学科信息资源在学术交流中产生，表现出很强的时效性、简洁性和非系统性，需要系统地组织以最大程度地发挥其学术促进功能。数字资源聚合通过强化语义和发现关联构建内容相互关联、多维度、多层次的资源体系，形成集概念主题、学科内容和科研对象于一体的知识网络[1]。本文采取数字资源聚合中聚类的知识再组织理念，在目标数字资源信息类别、体系结构和专业词表未知的情况下，针对学术资源网文献内部特征进行语义挖掘，采用关系分析和关联数据构建途径揭示学术资源网中的相似文献。

2 相关研究进展

关联文献发现主要以文献之间的相似性测度为基础，研究方向主要包括：以共词分析和向量空间模型等统计计算方法为基础的文档相似度计算方法；以知识体系语义理解为基础的语义相似度计算方法。

以统计计算为基础的文档相似度测算方法主要是针对构成某篇文献的主要词汇在其他文献中出现的频度进行计算，具有低成本、高效率的优势，突出表现在对文档集容量要求不高但测算精确度较高时，不需要领域词表的辅助也能顺利完成测算。Magerman 等[2]使用结合潜在语义分析与向量空间模型的文本挖掘技术评估专利与科技出版物之间的相似度，并通过人工

通讯作者：毕强，ORCID：0000-0001-7381-4986，E-mail：biqiang12345@163.com。

*本文系国家自然科学基金项目“语义网络环境下数字图书馆资源多维度聚合与可视化展示研究”(项目编号:71273111)和吉林大学高峰学科(群)建设项目的研究成果之一。

评定验证其效果。和晓萍等[3]提出基于预聚类的潜在语义文献检索算法，在潜在语义分析方法的基础上采用K-means(K均值)聚类算法，对待检索文档集进行预聚类寻找出各聚类簇的中心点，通过计算查询向量与各聚类簇中心点的相似度进行检索。Wang等[4]结合语义分析法与后向增殖神经网络对文本分类，借助潜在语义分析的统计推断能力形成概念向量空间，从而发现词汇间的重要关联并降低维度。Olmos等[5]采取结合语义描述动态模型、潜在语义空间降维和欧氏距离三种方法的人工分类器，以增强潜在语义分析方法的可靠性。

共词分析方法是在给定文献集合中词与词在同一篇文献中出现的频次的情况下，采用一定的统计方法对共现频次进行计算，得到词与词之间的关联强度，从而揭示信息在内容上的关联[6]。唐果媛等[7]通过分析数据集、主题演化阶段划分、选择和提取共词分析对象、构建共词矩阵与归一化、主题演化分析5步进行学科主题演化研究分析，并对每个步骤中研究人员使用的策略、分析手段和工具进行归纳总结。任建华等[8]提出一种利用关联规则强化的文档向量表示方法，在词条同现模型基础上，对文档中词条之间依托关联关系存在的潜在语义进行挖掘，在得到的文档向量中同时考察词条同现关系和词条间隐含关系，以提高聚类的准确性。

领域知识为基础的语义相似度测算方法往往使用较复杂的算法，需要借助完善的领域词表，成本较高，但可以赋予机器语义理解能力。经常被使用的领域词表包括“词网(WordNet)”和“知网(HowNet)”。如黄贤英等[9]提出一种词项语义维度映射方法，依据词频和HowNet词典完成词性向量权值映射，将短文本之间相似度运算转换为词性向量之间相似度运算。徐勇等[10]研究文献之间的相似程度度量问题，结合开放目录项目的目录系统构建文献关键词的泛化树结构，将关键词或其父子词语进行匹配反映两篇文献在研究领域视角上的相似性，加入共被引因素所代表的间接相似性，构成混合相似度识别语义相似的文献。如吴树芳等[11]提出基于术语间本体关联度的文档相关度计算方法，利用树状本体结构计算术语间基于本体的关联关系，通过术语组间本体关联度得到两组词语的本体关联关系，结合文档标引词权重计算两个文档的相关度。从本体角度将语义信息引入向量空间模型，提高文档相关度计算的准确性。

综合以上两个分支，能够发现潜在语义分析可以从词条、语句到篇章三个不同层次进行关联分析，有很高的自由度；共词分析以词对频次计量信息关联；语义关联则需要借助词表实现。本文采用灵活高效的潜在语义分析与拥有关系描述优势的关联数据相结合的方法，探讨通过统计分析方法建立关联数据，并实现相似文献发现的可行性和有效性。

3 基于关联数据的学术资源网相似文献聚合方法

学术资源网对文献的整理通常以所发布的文献的关键词标注为基准，但是关键词选择的标准因网站编辑与文献作者的不同而异，一些学术资源网的文献包含原文链接、原作者、编辑等简单的元数据，而部分文献则并不提供这些信息，在文献归类组织上表现出较大的随意性。因此本文通过潜在语义分析与向量空间模型对网络文档的内容进行语义相似性测算，将语义关联信息和文档元数据合并整理为关联数据，形成相似文献发现的基础。

3.1 潜在语义分析与向量空间模型

潜在语义分析(Latent Semantic Analysis, LSA)是一种全自动提取并推断语篇中词汇的预期语用关系的数学统计学技术，通过统计隐含在文本中词语的上下文使用模式，提取词汇之间潜在的语义结构[12]。它不使用词典等知识库工具，仅以原始文本按语法分解为词条作为输入数据。首先需要将文本表示成每一行是一个唯一的词汇，每一列是一个文档的矩阵。根据线性代数的奇异值分解(Singular Value Decomposition, SVD)方法能够分离出两个正交矩阵和一个对角矩阵乘积的原理，可以得到词条矩阵、由奇异值构成的秩矩阵和降维后的词条文档频率矩阵。将经奇异值分解所得行奇异向量和列奇异向量除以它们的维度所得的截断矩阵就构成了潜在语义空间(Latent Semantic Space)[13]。

文档相似度比较采用经典的向量空间模型(Vector Space Model, VSM)，将每一篇文献看作一个向量，一篇文献和另一篇文献的差异度就可以表示成这两个向量所构成的角度，用向量夹角的余弦表示[14]。同时，为

保证文档的相似关系不受高频词条的影响，采取词汇频率–逆向文档频率进行权重修正，获得文档相似度的准确结果。经过以上分析处理过程，从而获得建立关联数据所需的基础数据。

3.2　关联数据技术

关联数据是一组最佳实践的集合，它采用 RDF 数据模型，利用 URI(统一资源标识符)命名数据实体，发布和部署实例数据和类数据，从而可以通过 HTTP 协议揭示并获取这些数据，同时强调数据的相互关联、相互联系以及有益于人机理解的语境信息[15]。作为一种本体描述方法，其表述范畴包括概念、概念层次、属性、属性值类型、关系、关系定义域概念集以及关系值域概念集，并可以在此基础上添加规则或公理来表示模式层更复杂的约束关系[16]。关联数据实质是遵循：RDF 文档以统一资源定位符(URI)为名称；URI 必须符合超文本传输协议(HTTP)；URI 指向的信息必须以标准格式(RDF，SPARQL)提供；发布信息必须包含 URI 等关联数据原则[17]的 RDF 文档。

关联数据是构建网络知识组织体系的重要工具，它的对象标识与访问机制为跨区域信息资源聚合和信息资源追溯创造了良好的条件，同时也为各类对象实体以及所涉及的大量概念术语提供了规范控制。关联数据透过标准化的命名和指向，严格限定了数据的语义，也关联到其所链接的大量相关资源实体，这些关联数据的“属性”本身也是资源[18]。关联数据具有领域无关、机器可理解的特性，能够降低数据流动和转化过程经过人机交互或机器交互所产生的阻力，能够更好地携带语义数据，供用户访问和机器处理。

4　基于关联数据的学术资源网相似文献聚合框架与功能

学术资源网根据自身学科有选择地发布文献。受学科层次、覆盖范围和更新频率所限，网站只对这些文献进行简单分类，甚至仅统一收藏而不分类。用户在使用过程中只能靠逐个浏览来掌握自己需要的文献。鉴于此，如何有效实现学术资源网的相似文献聚合，为用户准确地展示检索的相似文档集就变得十分重要。

4.1　聚合框架

关联数据驱动的信息资源聚合框架主体包括三层结构，如图 1 所示：

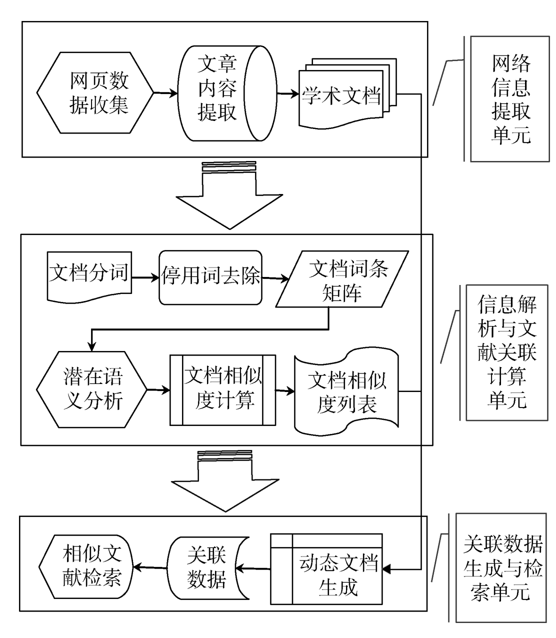


图 1　基于关联数据的学术网络信息相似资源聚合框架

实现数据由松散的网络信息向描述知识关联的关联数据的转化，最终实现关联文献的检索和推荐：

(1) 网络信息提取单元。实现学术资源网内部不同类型、不同时期、不同形式文献资源的提取，去除网页代码等无关内容。

(2) 信息解析与文献关联计算单元。对学术资源网文献中的元数据、学科知识等进一步萃取，向量空间化，计算各个文档之间的相似度，围绕每篇文档建立相似度图谱，为关联数据生成创造条件。

(3) 关联数据生成与检索单元。以网络信息提取单元获取的学术文档和学术信息解析与文献关联计量单元取得的文档相似度列表为基础，经动态文档生成系统生成完整的学科网络信息关联数据作为检索的核心资源，与用户信息检索行为无缝链接，提供快捷准确的关联文献检索服务，从而促进学术资源网络中知识的发现和利用。

4.2　聚合功能

学术资源网相似文献聚合是通过网络文献采集和预处理、文献相似度计算与筛选、关联数据生成和检索三个功能实现的。

(1) 网络文献采集和预处理

蕴藏于学术资源网中的文献大多嵌在网页文件之

中，需要经过采集和预处理构成基础语料库以备进一步的分析。其过程如图2所示:

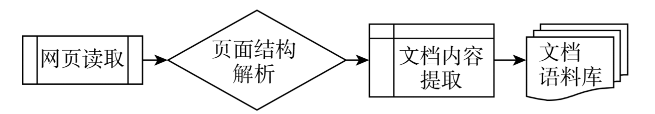


图2　网络文献采集和预处理过程

学术资源网文献采集包括网页读取与网页结构解析，网页读取实质就是通过爬虫、离线浏览器或文档传输(FTP)工具将目标网页从网络保存到本地的过程。为了在浏览器中按照预设的样式显示文档，文档内容和格式都通过一整套标记标签描述，每一组标签构成一个节点。提取网络文档内容时，需要首先对页面框架进行解析，从而准确选中包含文档内容的节点。预处理则是对采集到的网页节点内容进行去除冗余的页面代码、格式符号等清洗工作，随后将整理所得的内容按文档为单位分别保存，供分析使用。

(2) 文献相似度计算与筛选

学术资源网相似文献发现依赖于文档之间相似度的计算和筛选。本文采取潜在语义分析对文献资源之间的语义关联进行计算，在网络信息提取单元的文档语料库基础上，构建潜在语义分析向量空间，计算总体文献相似度；通过文档集层次聚类的中心趋势度确定过滤文献相似度的阈值，从而过滤掉相似度较低的文献。形成新的相似文献列表后，将其写入关联数据中，实现基于关联数据的相似文献资源语义聚合。文献相似度计算与筛选流程如图3所示:

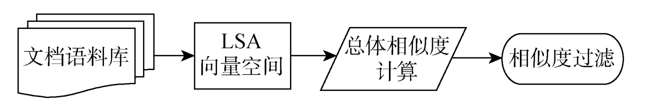


图3　文献相似度计算与筛选流程

文献相似度计算与筛选分为两个阶段:

①从文档语料库构建潜在语义分析向量空间，其主要功能是对由学术资源网中提取的学术文献的原始内容进行分词和去除停用词，建立文档词条矩阵并对词条频度和文档的关系进行权重计算。本文采用词汇频率–逆向文档频率(TF-IDF)计算文档相似度权重。词汇频率–逆向文档频率包含两部分功能：词汇频率(Term Frequency, TF)指词条在某篇文档中的出现次数，为防止某一关键词条在不同文档中出现频率差异导致权重波动较大，本文对词条对于单个文档的权重取对数，将其缩小至0到1的区间内；同时因文档词条矩阵为稀疏矩阵，因此在每个词汇频率取对数前加1，即 $TF(t,d)=\log(1+f_{t,d})$，防止最终的TF-IDF值过小造成比较的不便；逆向文档频率(Inverse Document Frequency, IDF)是文档集合中的文档总数和出现当前词项的文本数的比值，即 $IDF(t,D)=\log(N/|\{d\in D, t\in d\}|)$，它表示词项在文档集合中的普遍性。在判断文档相似性方面，如果一个词出现在文档集内越多的文献中，则它对文档的区分度越低，重要程度自然下降。TF-IDF就是TF和IDF的乘积，即 $TFIDF(t,d,D)=TF(t,d)\times IDF(t,D)$。文档词条矩阵建立在这些调整每篇文档与文档中的词汇关联程度权重之上。将所有待分析的n个文档顺次排列，按照全部文献中包含的m个词条出现频率组成一个矩阵，就构成 $n\times m$ 文档词条矩阵M。

②高相似度文档的提取，它依赖于整体相似度计算和相似文献过滤。本文采取潜在语义分析法对处理后的学术资源网文档的语义结构进行计算，将文档词条矩阵M经奇异值分解获得词条矩阵、奇异值矩阵与文档矩阵，即：$M=U\Sigma V^T$。奇异值矩阵与文档矩阵之积 ΣV^T 就是降维后的文档空间向量。文档间的相似度可以用降维后的两个文档在向量空间中的余弦值计算，该值越大，对应的两个文档相似度越高。在得到整体文档相似度之后，采取以文献层次聚类为基础的相似度阈值选取方法。聚类分析简称聚类，是把数据对象划分成子集的过程。每个子集是一个簇，簇中的对象彼此相似，而不同簇的对象不相似。由于它能够根据数据的相似性将大型数据集合划分成组，因此又称为数据分割[19]。层次聚类将初始种群的每个样本个体都单独作为一类，使用欧几里得距离评价各个类别之间的相似程度，并将最接近的类别进行合并，直到满足聚类需求[20]。通过对文档集进行层次聚类能够获得最大类簇容纳的文献数量，以此为依据对各篇文献的相似文献按照相似度从高到低的次序进行截取；将截取所得的相似度值集合求中位数即可进一步排除相似度较低的文献，得到合理的相似度阈值。

经过上述两个阶段，本文排除原始相似度矩阵中所有低相似度文献，得出记录不同类型、不同格式、不同著录规则的网络文献间的语义相似程度的文档相似度矩阵。

(3) 关联数据生成和检索

关联数据生成和检索模块将经过相似度过滤后的相似文献列表以动态文档技术写入关联数据内。该关联数据不必一次性容纳学术资源网中的所有文献，而是可以实时增补新文献信息。针对某种特定类型的文献资源，其表现的特有语义和语义关联可以动态地加入关联数据中，从而对核心元数据本体进行定制化扩展，生成针对某一学科门类的学科关联数据。

4.3　学术资源网相似资源表示

在关联数据中，学术资源之间的关联采用相似文献列表的形式展现是十分直观且方便检索的方法。笔者将某篇文献的相似文献作为关联数据中该文献属性的一个子类，以等同关系(similarAs)定义相似文献资源，根据文献相似度矩阵生成对应各个文献资源的相似文献列表。以博文《从统计学角度来看深度学习(2)：自动编码器和自由能》为例，其相似文献列表如图 4 所示：


```
<dcam:similarAs rdf:resource='http://cos.name/2015/06/a-statistical-view-of-deep-learning-iii-memory-and-kernels/'>
<dcam:similarAs rdf:resource='http://cos.name/2015/05/the-data-wisdom-for-data-science/'>
<dcam:similarAs rdf:resource='http://cos.name/2014/12/introduction-of-deep-learning/'>
<dcam:similarAs rdf:resource='http://cos.name/2015/01/talking-about-data-scientist/'>
<dcam:similarAs rdf:resource='http://cos.name/2015/02/the-application-of-statistics-in-love/'>
<dcam:similarAs rdf:resource='http://cos.name/2013/11/teaching-le-to-a-child/'>
<dcam:similerAs rdf:resource='http://cos.name/2015/01/the-material-of-data-science-for-the-vacation/'>
<dcam:similerAs rdf:resource='http://cos.name/2014/01/svm-series-maximum-margin-classifier/'>
<dcam:similerAs rdf:resource='http://cos.name/2015/03/using-r-to-search-for-your-partner/'>
<dcam:similerAs rdf:resource='http://cos.name/2014/03/svm-series-5-support-vector/'>
<dcam:similerAs rdf:resource='http://cos.name/2014/12/the-story-about-measure-theory/'>
<dcam:similerAs rdf:resource='http://cos.name/2014/02/svm-series-3-kernel/'>
<dcam:similerAs rdf:resource='http://cos.name/2013/10/simply-statistics-of-gmm/'>
<dcam:similerAs rdf:resource='http://cos.name/2015/04/interview-of-chutingjin/'>
<dcam:similerAs rdf:resource='http://cos.name/2012/02/what-is-the-stat-dept-25-years-from-now/'>
<dcam:similerAs rdf:resource='http://cos.name/2013/08/causality6-instrumental-variable/'>
<dcam:similerAs rdf:resource='http://cos.name/2014/05/svm-series-add-2-kernel-ii/'>
<dcam:similerAs rdf:resource='http://cos.name/2014/03/svm-series-add-1-duality/'>
<dcam:similerAs rdf:resource='http://cos.name/2013/05/relationship-big-data-statistics/'>
<dcam:similerAs rdf:resource='http://cos.name/2014/01/svm-series-2-support-vector/'>
```


图 4　关联数据中的相似文献列表样例

以关联数据的形式发布学术资源网信息资源语义关联列表，可以直观地展示全学科文献关联图谱，从而使整个学科的学术文献都可以由关联数据作为起始节点轻松访问，并可以经由统一资源定位符访问外部相关资源，自由地在不同数据集中进行切换。由于关联文献已经按关联度高低排序，因而能够通过简单的查询有效地揭示资源间的相互关系。此外，还能够实现语义检索等语义互操作。

5　实证分析

以真实的学术资源网数据为例，使用本文提出的基于本体与关联数据的学术资源网络相似文献聚合框架为基础，以 R 语言[21]为工具构建一个演示性的学术资源网相似文献聚合系统，实现学术资源网相似文献的推荐。

5.1　数据与预处理

选取学术资源网“统计之都”发布的编辑推荐文献共 78 篇。全程采用 XML 程辑包[22]对“推荐文献”栏目页面中所含文献链接进行提取；通过这些链接预读取推荐文献全文页面；将包含文献内容的节点提取出来，每一篇文献单独组织成一个文档，并借用“统计之都”网站的链接特征，以该文献统一资源定位符的最后一节对其命名。使用基于隐马尔科夫模型的中国科学院计算技术研究所的 ICTCLAS 分词软件①的 Rwordseg 程辑包[23]统一读取这些待分析的文档对象并实现分词和去除停用词，以降低无实际意义的词造成的系统资源消耗。

5.2　相似文档聚类与发现

本文调用 lsa 程辑包[24]，将每一个按词条分解的文档作为一个单独的向量读入，形成原始语料库。由于中西文差异，笔者将最短词长调整为 1，以便在尽可能保留中西文词条的前提下对语料库进行清洗。将清洗后的语料库转换为文本矩阵，此时该文本矩阵即是由分解的词条组成的文档向量集。在计算词频时，采取词汇频率–逆向文档频率(TF-IDF)作为平衡高频词权重的方法建立文档词条矩阵，并对文档词条矩阵进行奇异值分解和可视化，结果如图 5 所示。

图 5 即经过奇异值分解降维后的文档词条矩阵，它清晰地展示了学术资源网“统计之都”中发布的推荐文献所构成的各个文档向量的关联情况。图中左侧集中的点簇展现了“统计之都”推荐文献的语义相似

①http://ictclas.nlpir.org.

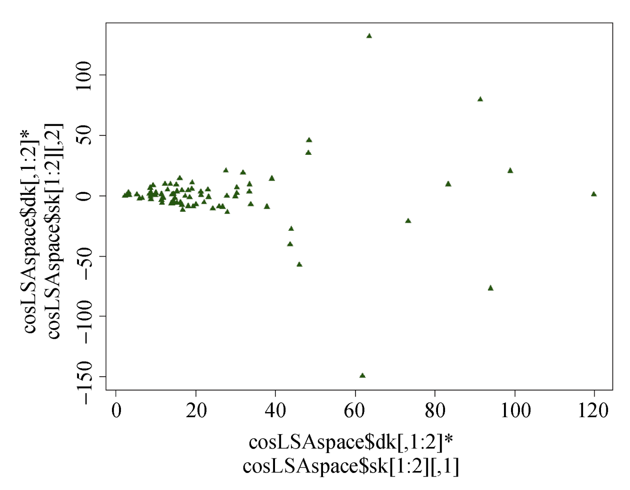


图 5 “统计之都”文档潜在语义空间双奇异值分解图

度关系，且大多数文献都聚集在左侧区域内，仅在右侧散布着8个离群点，表示这8篇文献与其他文献的关联并不紧密。这一现象充分体现了学术资源网主推的文献具有较高的学科主题一致性，但是也可能存在与其他文献关联较弱的灰色文献。在此基础上，对降维后的文档词条矩阵求余弦相似度，即可得出整体文档相似关系矩阵。

为准确检索出最相似的文档，使用基于层次聚类的相似度阈值求解方法。调用 R 平台的 proxy 程辑包[25]分两步实现层次聚类：求得文档间的欧氏距离；采用离差平方和进行层次聚类。离差平方和法对文档聚类的判断基于方差分析思想，如果分类合理，则同类文档之间的离差平方和应当较小，不同类间的离差平方和应当较大。经过计算与可视化，得到“统计之都”推荐文档层次聚类，如图 6 所示：

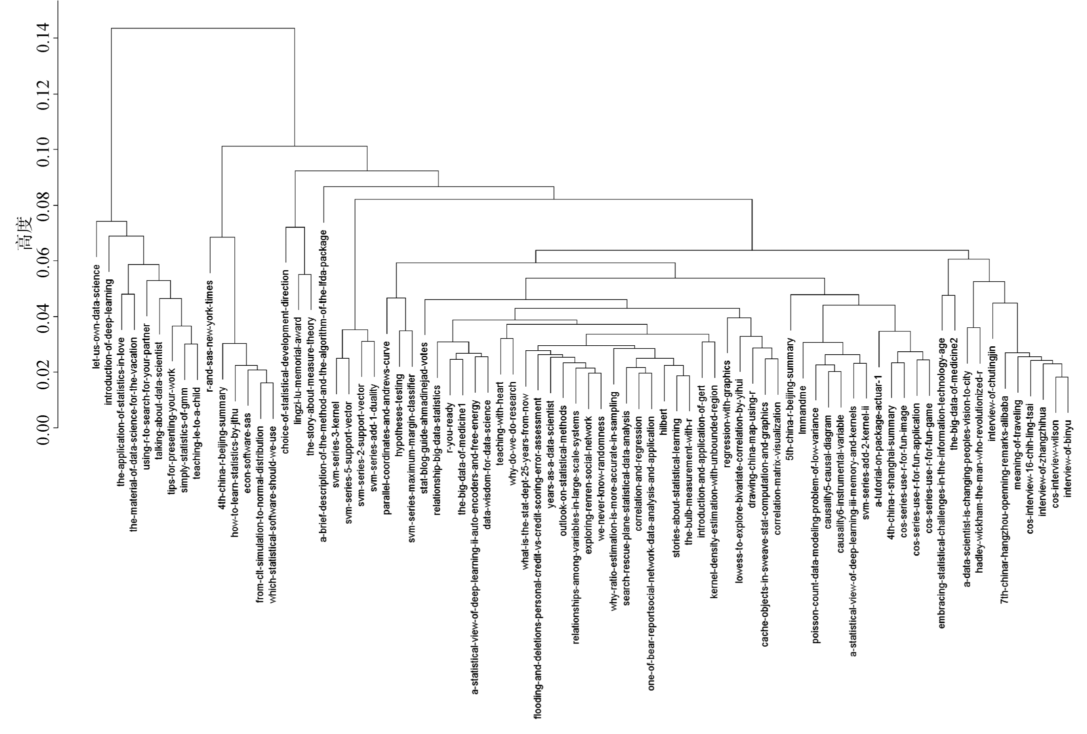


图 6 “统计之都”文档的层次聚类分析

图 6 显示了所有推荐文档的聚类情况，当高度选择 0.04 时，层次聚类获得的类簇包含 4 至 29 篇文献不等。为完整保留高相似度文献关联，以最大类簇包含的文档数作为提取相似文档的初始值，再分别从 29 个文档中截取高相似度的数据，将这些符合条件的相似度数据汇总并求取中位数，得到相似度阈值 0.6321。以此阈值为基准，对文档相似度矩阵中的数据进行再次筛选，保留所有高于阈值的相似度值。若某篇文档与其他文档相似度均低于阈值，则仅选取与其相似度最高的一篇文档作为其相似文献。在获得文献之间的关联度基础上，对每一篇文献按照相关度由高到低的次序生成全文档集的相关文献列表。

在相关文献列表生成后，利用动态文档转换工具 rmarkdown[26] 将文档元数据和相关文档列表嵌入

RDF/XML 编码段内形成完整的关联数据源码动态文档，经过简单的格式转换即可获得记录相似文档的关联数据。借助社会网络分析与可视化工具 iGraph[27]，笔者对记录在关联数据中的文档相似关系加以可视化展示，如图 7 所示：

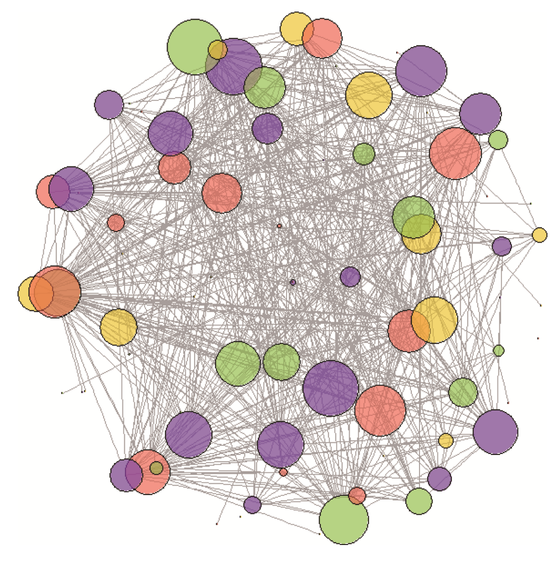


图 7　关联数据中记录的“统计之都”推荐文档的相似关系

图 7 中节点的大小与该文档和其他文档的相似度相关，相似文档越多节点越大，网络图边缘散落的单连线小节点就是那些与其他文档相似度都较低(不及阈值)的少量文献。本文使用的相似度处理方法最大限度地保留了这些文档与其他文档的相似关系，使之在相似文献检索过程中能够发现。只需在网络上发布关联数据，运用“跟着感觉走”原则(Follow Your Nose Principle)只要确定一个指向某些 RDF 的 URI，就可以引用这个 URI 加载相应文档[28]，迅速获得“统计之都”中某篇文献的相关文献检索结果。

6　结　语

关联数据实现相似文献推荐具有较高的灵活性：关联数据是文档分析过程的成果，用户检索针对关联数据进行，只需一次简单检索或推理即可得出结果，无须将查询与每篇文档依次比对，提高检索效率；相关检索结果直接由关联数据给出，并以 URI 形式向用户提供，直接点击即可获得，符合广大用户的使用习惯，简单方便。由于分析过程的独立性，在文档发生变化时不必中断服务。由系统在后台分析完成并生成新的关联数据后替换原有旧关联数据，可以完全不干扰用户的使用。

本文采用潜在语义分析方法对网络文档包含的语义进行计算，以得到的相似度矩阵作为生成关联数据的基础，展示此类方法在相似文献发现中的有效性。文章对关联数据的构建比较简单，主要表征了文档间的相似性，以及利用相似关联关系对新的相似文献进行发现。并未对文档所涉及的学科知识进行归类和关联规则发掘。以客观知识体系和知识结构为基础的相似文献发现应更能体现学科知识发展脉络和相关文献的关联程度。后续研究中笔者将引入机器学习等方法对学术资源网的文献内容和知识进行深度聚合。

参考文献：

[1] 张云中．从整合到聚合：国内数字资源再组织模式的变革[J]．数字图书馆论坛，2014(6)：16-20．(Zhang Yunzhong. From Integration to Aggregation: The Change of Digital Resources Re-organization Pattern in China [J]. Digital Library Forum, 2014(6): 16-20.)

[2] Magerman T, Van Looy B, Song X. Exploring the Feasibility and Accuracy of Latent Semantic Analysis Based Text Mining Techniques to Detect Similarity Between Patent Documents and Scientific Publications [J]. Scientometrics, 2010, 82(2): 289-306.

[3] 和晓萍，李迪，王米利，等．基于预聚类的潜在语义分析模型文献检索研究[J]．云南民族大学学报：自然科学版，2015, 24(3): 257-260. (He Xiaoping, Li Di, Wang Mili, et al. A New Pre-Clustering-based Latent Semantic Analysis Algorithm for Document Retrieval[J]. Journal of Yunnan Nationalities University: Natural Sciences Edition, 2015, 24(3): 257-260.)

[4] Wang W, Yu B. Text Categorization Based on Combination of Modified back Propagation Neural Network and Latent Semantic Analysis [J]. Neural Computing & Application, 2009, 18(8): 875-881.

[5] Olmos R, León J A, Jorge-Botana G, et al. New Algorithms Assessing Short Summaries in Expository Texts Using Latent Semantic Analysis [J]. Behavior Research Methods, 2009, 41(3): 944-950.

[6] Law J, Bauin S, Courtial J P, et al. Policy and the Mapping of Scientific Change: A Co-word Analysis of Research into Environmental Acidification [J]. Scientometrics, 1988, 14(3):

251-264.

[7] 唐果媛，张薇．基于共词分析法的学科主题演化研究进展与分析[J]．图书情报工作，2015，59(5)：128-136．(Tang Guoyuan, Zhang Wei. Development and Analysis of Subject Theme Evolution Based on Co-word Analysis Method [J]. Library and Information Service, 2015, 59(5): 128-136.)

[8] 任建华，沈炎彬，孟祥福，等．基于词条之间关联关系的文档聚类[J/OL]．[2014-12-11]．计算机工程与应用．http://www.cnki.net/kcms/detail/11.2127.TP.20141211.1528.053.html. (Ren Jianhua, Shen Yanbin, Meng Xiangfu, et al. Document Clustering Based on Association Relations Between Terms [J/OL]. [2014-12-11]. Computer Engineering and Applications. http://www.cnki.net/kcms/detail/11.2127.TP.20141211.1528.053.html.)

[9] 黄贤英，张金鹏，刘英涛，等．基于词项语义映射的短文本相似度算法[J]．计算机工程与设计，2015，36(6)：1514-1518，1534．(Huang Xianying, Zhang Jinpeng, Liu Yingtao, et al. Short Text Similarity Algorithm Based on Term Mapping with Semantic[J]. Computer Engineering and Design, 2015, 36(6): 1514-1518, 1534.)

[10] 徐勇，陈建国，胡凌云，等．基于泛化语义相似的科技文献混合推荐算法[J]．情报理论与实践，2013，36(2)：96-99，103．(Xu Yong, Chen Jianguo, Hu Lingyun, et al. S&T Literature Hybrid Recommendation Algorithm Based on Generalized Semantic Similarity [J]. Information Studies: Theory & Application, 2013, 36(2): 96-99, 103.)

[11] 吴树芳，刘畅，徐建民．基于术语间本体关联度的文档相关度研究[J]．现代情报，2014，34(9)：56-59，176.(Wu Shufang, Liu Chang, Xu Jianmin. Research on Document Relevancy Based on Ontology Term Relations [J]. Journal of Modern Information, 2014, 34(9): 56-59, 176.)

[12] Steyvers M, Griffith T. Probabilistic Topic Models[A].// Latent Semantic Analysis: A Road to Meaning [M]. Laurence Erlbaum, 2006.

[13] Landauer T K, Foltz P W, Laham D. An Introduction to Latent Semantic Analysis [J]. Discourse Processes, 1998, 25(2-3): 259-284.

[14] Leydesdorff L. Similarity Measures, Author Cocitation Analysis, and Information Theory [J]. Journal of the American Society for Information Science & Technology (JASIST), 2005, 56(7): 769-772.

[15] Structured Dynamic. Linked Data FAQ [EB/OL]. [2014-07-18]. http://structureddynamics.com/linked_data.html.

[16] 王昊奋．大规模知识图谱技术[J]．中国计算机学会通讯，2014，10(3)：64-68．(Wang Haofen. Large-scale Knowledge Graph Technology [J]. Communications of the CCF, 2014, 10(3): 64-68.)

[17] Berners-Lee T. Linked Data-Design Issues [EB/OL]. [2009-06-18]. http://www.w3.org/DesignIssues/LinkedData.html.

[18] 刘炜．关联数据:概念、技术及应用展望[J]．大学图书馆学报，2011，29(2)：5-12．(Liu Wei. Overview on Linked Data: Concept, Technology and Implementation[J]. Journal of Academic Libraries, 2011, 29(2): 5-12.)

[19] Han J, Kamber M, Pei J．数据挖掘：概念与技术[M]．范明，孟小峰译．第 3 版．北京：机械工业出版社，2012: 288-289. (Han J, Kamber M, Pei J. Data Mining: Concept and Techniques[M]. Translated by Fan Ming, Meng Xiaofeng. The 3rd Edition. Beijing: China Machine Press, 2012: 288-289.)

[20] Tang X, Zhu P. Hierarchical Clustering Problems and Analysis of Fuzzy Proximity Relation on Granular Space[J]. IEEE Transactions on Fuzzy Systems, 2013, 21(5): 814-824.

[21] The R Project for Statistical Computing [EB/OL]. [2015-07-10]. https://www.R-project.org/.

[22] XML: Tools for Parsing and Generating XML Within R and S-Plus [EB/OL]. [2015-06-30]. http://CRAN.R-project.org/package=XML.

[23] Rwordseg: Chinese Word Segmentation[EB/OL]. [2013-12-15]. http://R-Forge.R-project.org/projects/rweibo/.

[24] lsa: Latent Semantic Analysis[EB/OL]. [2015-05-27]. http://CRAN.R-project.org/package=lsa.

[25] proxy: Distance and Similarity Measures[EB/OL]. [2015-07-08]. http://CRAN.R-project.org/package=proxy.

[26] rmarkdown: Dynamic Documents for R [EB/OL]. [2015-06-13]. http://CRAN.R-project.org/package=rmarkdown.

[27] Csardi G, Nepusz T. The iGraph Software Package for Complex Network Research [C]. In: Proceedings of Inter-Journal, Complex Systems Cambridge, MA USA. 2006: 1695.

[28] Antoniou G, Groth P, Hoekstra R，等．语义网基础教程[M]．胡伟，程龚，黄智生译．第 3 版．北京：机械工业出版社，2014. (Antoniou G, Groth P, Hoekstra R, et al. A Semantic Web Primer [M]. Translated by Hu Wei, Cheng Gong, Huang Zhisheng. The 3rd Edition. Beijing: China Machine Press, 2014.)

作者贡献声明:

毕强：提出研究思路，设计研究方案；

赵夷平：进行数据采集、实验与论文起草；

毕强，赵夷平：论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 赵夷平, 毕强. category_url. “统计之都”推荐文章栏目包含的网页链接.

[2] 赵夷平, 毕强. article_url. 由“统计之都”推荐文章栏目中提取的全部推荐文章链接.

[3] 赵夷平, 毕强. ctl.csv. “统计之都”推荐文章链接与标题.

[4] 赵夷平, 毕强. cate80.rar. 经过分词后的“统计之都”推荐文章.

[5] 赵夷平, 毕强. ccsm.csv. 原始文献相似度矩阵.

[6] 赵夷平, 毕强. cosArtSim.csv. 去除低相似度的文献相似度矩阵.

[7] 赵夷平, 毕强. cossim.rdf. “统计之都”相似推荐文章关联数据.

收稿日期: 2015-08-13
收修改稿日期: 2015-10-04

Using Linked Data to Retrieve Similar Documents from the Academic Resource Websites

Zhao Yiping　Bi Qiang
(School of Management, Jilin University, Changchun 130022, China)

Abstract: [**Objective**] This paper studied the linked data from the Web, which is machine-readable, semantically meaningful and relationally descriptive. We examined these data's effectiveness to improve the information organization of the academic resource websites (ARWs), with the purpose of retrieving more similar documents. [**Methods**] We first calculated the similarity of documents published in the ARWs with the help of the Latent Semantic Analysis (LSA) method. Then, chose documents with high similarities by the Hierarchical Cluster method, and created a document relation matrix. Finally, we used the dynamic document technology to generate a linked data index to search the ARWs. [**Results**] We built a preliminary ARWs linked data index, which helped us find similar documents more effectively from the ARWs. [**Limitations**] We investigated the similar documents retrieval technology from the perspective of statistical analysis. Therefore, further research is needed to locate similar documents from various subject areas with the support of deep learning technology. [**Conclusions**] We computed documents' similarity using LSA method to discover related documents of specific articles. The linked data could help us find more similar documents, while reducing the waiting time for similarity calculation.

Keywords: Linked data　Latent Semantic Analysis(LSA)　Academic Resource Websites(ARWs)　Similarity